面向汉维机器翻译的双语关联度优化模型 *

潘一荣 [1, 2, 3]，李　晓 [1, 3+]，杨雅婷 [1, 3]，董　瑞 [1, 3]

(1. 中国科学院 新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院大学，北京 100049；3. 新疆民族语音语言信息处理实验室，乌鲁木齐 830011)

摘　要：针对汉语-维吾尔语的统计机器翻译系统中存在的语义无关性问题，提出基于神经网络机器翻译方法的双语关联度优化模型。该模型利用注意力机制捕获词对齐信息，引入双语短语间的语义相关性和内部词汇匹配度，预测双语短语的生成概率将其作为双语关联度，以优化统计翻译模型中的短语翻译得分。在第十一届全国机器翻译研讨会(CWMT 2015)汉维公开机器翻译数据集上的实验结果表明，与基线系统相比，在使用较小规模的训练数据和词汇表的条件下，所提方法可以有效地同时提高短语级别和句子级别的机器翻译任务性能，分别获得最高 2.49 和 0.59 的 BLEU 值提升。

关键词：维吾尔语；神经网络机器翻译；注意力机制；词对齐；生成概率；

中图分类号：H085　　doi: 10.3969/j.issn.1001-3695.2018.08.0625

Bilingual relatedness optimization model for chinese-uyghur machine translation

Pan Yirong[1, 2, 3], Li Xiao[1, 3+], Yang Yating[1, 3], Dong Rui[1, 3]

(1. *Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi* 830011, *China*; 2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*; 3. *Xinjiang Laboratory of Minority Speech & Language Information Processing, Urumqi* 830011, *China*)

Abstract: Focused on the issue of semantic independence in Chinese-Uyghur statistical machine translation system, this paper proposed a bilingual relatedness optimization model based on neural machine translation method. The model utilized the attention mechanism to capture word alignment information as well as introduced bilingual phrase semantic relevance and inner word correlation to predict the conditional probability of bilingual phrase pair. And then took the probability as bilingual relatedness to optimize the phrase translation scores in statistical translation model. Experimental results on the 11th China Workshop on Machine Translation (CWMT 2015) Chinese-Uyghur public machine translation datasets show that the proposed approach can achieve obvious improvements both in the phrase-level and the sentence-level machine translation tasks, which outperforms the baseline system with a relative small-scale training data and vocabulary. The highest BLEU point gains are 2.49 and 0.59 respectively.

Key words: Uyghur; neural machine translation; attention mechanism; word alignment; conditional probability;

0　引言

在基于短语的统计机器翻译(statistical machine translation, SMT) [1]系统中，翻译模型对从平行语料库中抽取的双语短语进行建模，主要包括短语翻译概率、词汇化权重等参数，这些参数作为特征函数并结合对数线性方法，以此训练机器翻译系统，从而获取最优权重分布，在解码时以搜索最有可能的翻译选项，实现双语转换过程。虽然各种机器翻译方法在近年来取得了巨大进步，机器翻译质量也在不断提高，但是译文结果中存在的词汇翻译错误、语义内容无关等问题仍有待提升。

SMT 基于统计学方法构建短语翻译得分，仅考虑双语短语的共现频率，在一定程度上忽略语义相关性；同时由于词对齐结果来源于统计对齐模型[2]，并使用最大似然估计(maximum likelihood estimation, MLE)方法进行学习，所以存在缺失、冗余、错误等信息，在评估统计翻译模型中的词汇化权重时面临数据稀疏性等问题，降低其准确性，进而影响机器翻译质量。针对上述问题，本文应用神经网络机器翻译(neural machine translation, NMT)方法，并基于注意力机制捕获词对齐信息，引入双语短语的语义相关性和内部词汇匹配度，预测双语短语的关联度得分，以优化统计翻译模型的短语翻译概率，并在实验中证明了其有效性。

1　相关工作

维吾尔语属于小语种，且词形结构复杂，目前针对汉语-维吾尔语的统计机器翻译方法主要集中在两方面：一方面是平行语料库构建，例如彭飞等人[3]利用汉维双语语句的空间向量表示，通过源语句与目标语句间的相似度进行平行语料抽取，以保证平行句对在语义内容上的相关性；另一方面是维吾尔语语法以及形态分析，例如米莉万等人[4]将维语词汇

收稿日期：2018-08-22；**修回日期：**2018-10-23　**基金项目：**国家自然科学基金资助项目（U1703133）；中科院西部之光项目（2017-XBQNXZ-A-005）；中国科学院青年创新促进会的资助项目（2017472）；新疆维吾尔自治区重大科技专项项目（2016A03007-3）；新疆维吾尔自治区高层次人才引进工程项目（Y839031201）

作者简介：潘一荣（1992-），女，天津人，博士研究生，主要的主要研究方向为自然语言处理、机器翻译；李晓（1957-），男（通信作者），工学研究员，硕士，博导，主要研究方向为多语种信息处理、信息系统研究与开发（xiaoli@ms.xjb.ac.cn）；杨雅婷（1985-），女，副研究员，硕导，博士，主要研究方向为多语种信息处理技术；董瑞（1985-），男，助理研究员，博士研究生，主要研究方向为多语种信息处理.

的词干和词缀作为基本翻译单位，提出基于有向图的维吾尔语词干-词缀语言模型，利用维吾尔语的黏着语特性进行机器翻译实验。此外，潘一荣等人[5]利用深度学习技术对汉维短语的语义特征进行分析，通过循环神经网络(recurrent neural networks, RNN)学习调序信息并重构汉维调序模型，赋予调序规则更加合理的调序方向以及概率分布。上述方法重点在于对单语的语言特性(维语词干、词缀、形态等)以及双语外部对应关系(汉维平行句对、调序方向等)进行建模，缺乏关于双语对齐短语的语义相关性和内部词汇匹配度的研究和分析，故汉维统计翻译模型中存在语义无关性问题。由于短语翻译概率分布值并不合理，所以无法正确评估双语短语在语义及词汇上的关联度，在统计机器翻译研究中仍有提升空间。

随着深度学习技术在 SMT 领域中的应用和发展，许多研究工作利用神经网络方法改进统计翻译模型，并取得一定的效果。Schwenk [6]提出基于短语的连续空间翻译模型，该模型利用短语的向量表示来预测短语的翻译概率；Son 等人[7]提出分层结构的神经网络翻译模型，对翻译单元的连续空间向量表示及相关参数进行联合评估；Zou 等人[8]提出基于双语的词向量表示模型，对词汇间的语义相似度进行计算，并且将得分作为额外特征加入翻译系统的训练过程中；Cho 等人[9]提出基于编码器—解码器的短语表示模型，该模型利用 RNN 进行训练以最大化对齐短语的条件概率，并评估翻译模型中双语短语的生成概率得分。

本文参考(Cho et al., 2014)的工作，对统计翻译模型中的短语翻译概率进行重新评估，提出基于神经网络方法的双语关联度优化模型(neural-based bilingual relatedness optimization model, NBROM)。不同于上述研究思路，首先本模型基于(Bahdanau et al., 2014)框架[10]，利用注意力机制捕获双语短语间的词对齐信息，引入短语语义相关性和内部词汇匹配度，优化短语翻译概率；其次在训练该模型时，对于维吾尔语中的未登录词(out-of-vocabulary, OOV)问题，本文使用三种模型进行 OOV 词汇的生成概率预测，分别为 Unk 模型、MultiClass 模型和字节对编码(byte pair encoding, BPE)模型[11]，对不同的 OOV 词汇赋予相应权重。在第十一届全国机器翻译研讨会汉维数据集上的实验结果表明，在使用较小规模的训练数据和词汇表的条件下，与基线系统相比，本文提出的方法可以同时提高短语级别和句子级别的机器翻译任务性能，分别获得最高 2.49 和 0.59 的 BLEU 值提升，验证了本方法的有效性。

2 统计机器翻译系统

给定一个源语言语句 $\boldsymbol{f}$，SMT 的目标是找出相应的最优目标语言翻译结果 $\boldsymbol{e}$，使得条件概率最大化，如式(1)所示。

$$p(\boldsymbol{e}|\boldsymbol{f}) \propto p(\boldsymbol{f}|\boldsymbol{e}) \cdot p(\boldsymbol{e}) \tag{1}$$

其中：$p(\boldsymbol{f}|\boldsymbol{e})$ 为翻译模型；$p(\boldsymbol{e})$ 为语言模型。一般来说，SMT 将多种特征函数与其对应权重共同加入至对数线性框架中，并使用 $\log p(\boldsymbol{e}|\boldsymbol{f})$ 进行建模，如式(2)所示。

$$\log p(\boldsymbol{e}|\boldsymbol{f}) \propto \log p(\boldsymbol{f}|\boldsymbol{e}) + \log p(\boldsymbol{e}) \tag{2}$$

$$\propto \sum_{n=1}^{N} w_n \cdot f_n(\boldsymbol{f}, \tilde{\boldsymbol{e}}) + \log Z(e)$$

其中：f_n 和 w_n 分别为第 n 个特征函数及相应权重值；$Z(e)$ 为归一化常数项。基于该框架，翻译模型因子化为特征函数的加权总和。SMT 通过在开发集上优化权重参数 w_n，最大化翻译性能指标 BLEU 值[12]，并使用这些参数在解码过程中搜索最优候选短语翻译。

统计翻译模型主要对短语翻译概率和词汇化权重进行建模。短语翻译概率由统计方法进行计算，如式(3)所示。

$$Pr(\bar{e}|\bar{f}) = \frac{count(\bar{f}, \tilde{e})}{\sum_{\bar{e}_i} count(\bar{f}, \bar{\tilde{e}}_i)} \tag{3}$$

其中：$\bar{f}$ 和 $\bar{e}$ 分别表示源语言和目标语言的对齐短语；$count(\bar{f}, \tilde{e})$ 表示两者在较大规模平行句对中的共现频率。翻译模型使用该值作为此对齐短语的翻译概率。词汇化权重得分以词对齐结果为基准，将源语言和目标语言短语划分为词汇，用于评估双语词汇间的匹配程度，如式(4)所示。

$$lex(\bar{e}|\bar{f}, \tilde{a}) = \prod_{i=1}^{I} \frac{1}{|\{j \mid (i, \tilde{j}) \in a\}|} \sum_{\forall (i, \tilde{j}) \in a} (e_i|f_j) \tag{4}$$

$$(e_i|f_j) = \tilde{\ } \frac{count(f_j, \tilde{e}_i)}{\sum_{e_i} count(f_j, \tilde{e}_i)}$$

其中：$count(f_j, \tilde{e}_i)$ 表示词对 $(f_j, \tilde{e}_i)$ 在大规模平行句对中的共现频率。对于目标端短语 $\bar{e}$，翻译模型对其中的全部词汇按序进行遍历并连乘概率值，将此值作为词汇化权重得分。词对齐由 GIZA++[13]工具进行获取，该工具基于 EM 算法，用于评估词对的统计对齐概率，认为概率最大值对应的双语词汇在平行语料中对齐。由于上述两者基于统计方法获取，翻译模型中存在语义无关性问题。

汉维双语对齐短语实例（维吾尔语从右至左书写）如图 1 所示。对于源语言短语【为 人民 群众 服务】，统计翻译模型保留从训练语料抽取的对齐目标短语【ئامما ئۈچۈن خىزمەت】，并且赋予该对齐短语相应的词对齐信息。由此可以看出，源短语中的词汇【服务】对应于两个目标词汇【ئۈچۈن】和【خىزمەت】，然而只有【خىزمەت】符合语义对齐要求，【ئۈچۈن】对齐信息存在错误；同时词汇【为】和【人民】缺少对齐目标词汇。由于词汇化权重得分为对齐词汇的翻译概率乘积，若词对齐信息存在缺失、冗余、错误等问题，词汇化权重将无法正确评估双语短语中词汇间的匹配程度，从而降低统计翻译模型准确性以及 SMT 系统性能。

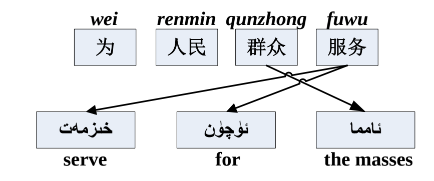


图 1 汉维双语对齐短语实例

Fig. 1 Example of Chinese-Uyghur bilingual aligned phrase

3 双语关联度优化模型

3.1 模型概述

基于深度学习技术和 NMT 方法，本文中模型首先使用编码器将源语言短语编码为固定维度的特征向量；然后使用解码器将该向量解码为不定长度的目标短语，并引入注意力机制，用于捕获双语短语中的语义信息和对齐词汇，以弥补统计翻译模型中存在的语义无关性和词对齐错误等不足。

编码器是一个双向循环神经网络 (bidirectional recurrent neural network, BiRNN) [14]，用于双向处理输入序列 $\boldsymbol{x} = (x_1, \ldots, x_t)$，分别更新正向隐藏状态($h(\rightarrow)_1, \ldots, h(\rightarrow)_t$)以及逆向隐藏状态($h(\leftarrow)_1, \ldots, h(\leftarrow)_t$)，如式(5)(6)所示。

$h(\rightarrow)$ k=f($h(\rightarrow)$ k-1,xk) xk 由 x1 到 xt 正序进行遍历 (5)

$\overleftarrow{h}_k = f(\overleftarrow{h}_{k-1}, x_k)$　　xk 由 xt 到 x1 逆序进行遍历　(6)

其中：f 为非线性激活函数 *tanh*。对于源短语中的每个单词 x_k，使用 $h_k = [\overrightarrow{h}_k^T, \overleftarrow{h}_k^T]$进行标注，引入其周围词汇信息。

解码器是一个单层 RNN，给定前一时刻的预测单词 y_{i-1}、当前时刻的 RNN 隐藏状态 s_i 和上下文向量 c_i，预测当前时刻的输出单词 y_i，其中 y_i 和 s_i 都依赖于 y_{i-1} 和 c_i，如式(7)所示。

$$p(y_i \mid y_1,\ldots,y_{i-1},\vec{x}) = g(y_{i-1},\vec{s}_i,\vec{c}_i) \tag{7}$$

$$s_i = f(s_{i-1},\vec{y}_{i-1},\vec{c}_i)$$

$$c_i = \sum_{j=1}^{t} \alpha_{ij} h_j$$

其中：g 为非线性激活函数 *softmax*；c_i 为 h_j 的加权总和；α_{ij} 为对齐权重，用于评估对齐模型 e_{ij} 中的单词 x_j 与 y_i 间的匹配程度，如式(8)所示。

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{q=1}^{t} \exp(e_{iq})} \tag{8}$$

$$e_{ij} = a(s_{i-1},\vec{h}_j)$$

$$a(s_{i-1},\vec{h}_j) = tanh(W_a s_{i-1} + U_a h_j)$$

其中：W_a 和 U_a 是神经网络参数。通过训练编码器和解码器，以最大化对数条件概率分值，预测最有可能的目标短语，如式(9)所示。

$$max_\theta \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y_n \mid x_n) \tag{9}$$

其中：θ 表示模型参数集合，(x_n，y_n)为训练数据集合中的对齐短语。以汉维双语短语生成过程为例，本文模型框架如图 2 所示。该模型将源语言短语映射为连续空间向量表示，用于获取其中的语义信息，同时利用注意力机制，对源短语中的各个词汇赋予不同的权重，以表征该词汇的重要性。利用此模型，可以在给定源语言短语的条件下，预测最有可能与之对应的目标语言短语；并可以预测双语短语的关联度得分，重新评估统计翻译模型中的词汇化权重，同时提高短语级别和句子级别的机器翻译质量，具体如下文所述。

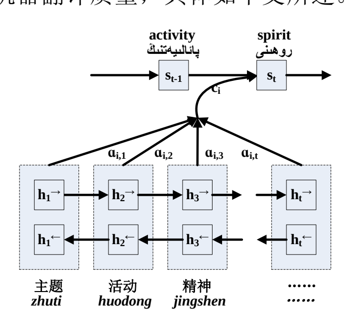


图 2　汉维神经网络双语关联度优化模型框架

Fig. 2　Framework of Chinese-Uyghur neural-based bilingual relatedness optimization model

3.2　目标短语预测

利用上述模型，可在给定源语言短语 s 的条件下，预测目标短语 e^*，使得式(9)的得分最大化，如式(10)所示。

$$\boldsymbol{e}^* = \arg\max_{e} \log p(\boldsymbol{e} \mid \boldsymbol{s}) \tag{10}$$

$$= \arg\max_{e} \sum_{t=1}^{I} \log p(\boldsymbol{e}_t \mid \boldsymbol{s})$$

$$= \arg\max_{e} \sum_{t=1}^{I} \log g(\boldsymbol{e}_{t-1},\vec{s}_t,\vec{c}_t)$$

其中：$\boldsymbol{e}_t$ 为预测短语中第 t 个词汇；I 为 $\boldsymbol{e}^*$ 中包含的词汇个数。

3.3　双语关联度得分

该模型同样可对统计翻译模型中的词汇化权重值进行重新评估，考虑双语短语的语义相关性和内部词汇匹配度，预测合理的双语关联度得分。给定源短语 $\bar{f}$，预测目标短语 $\bar{\boldsymbol{e}} = (\bar{e}_1,\ldots,\bar{e}_t)$ 的关联度得分，如式(11)所示。

$$pro(\bar{e} \mid \bar{f}) = \sum_{t=1}^{I} \log p(\bar{e}_t \mid \bar{e}_{<t},\vec{f}) \tag{11}$$

$$= \sum_{t=1}^{I} \log g(\bar{e}_{t-1},\vec{s}_t,\vec{c}_t)$$

其中：$\bar{e}_t$ 为目标短语中第 t 个词汇；I 为目标短语 $\bar{e}$ 包含的词汇个数。

3.4　未登录词处理策略

由于在训练 NMT 系统时，考虑到时间和空间复杂度，只保留频率较高的前 K 个目标词汇，K 取值一般在 30*k* (Bahdanau et al., 2015)~80*k* (Sutskever et al., 2014) [15]之间，所以无法有效地预测频率较低的稀有词汇，从而降低 NMT 系统的翻译质量。针对上述问题，本文使用以下三种模型对 OOV 词汇进行生成概率预测，并且在实验中验证其各自的有效性和局限性。对于源语言所有的 OOV 词汇，本文统一使用[UNK]符号进行标注，并设置源语言词汇表大小为 50*k*，目标语言词汇表大小为 30*k*-50*k*。

3.4.1 Unk 模型

参考(Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015)的工作，该模型对所有的 OOV 词汇统一使用[UNK]符号进行标注，这是一种非常普遍的未登录词处理策略，在 NMT 系统训练时，所有 OOV 被赋予相同的权重。

3.4.2 MultiClass 模型

参考(Jean et al., 2015)的思路[16]，该模型对所有的 OOV 进行词汇类别分类，同一类的 OOV 词汇具有相同的权重。对于各个未登录词，该模型在预测该词汇的生成概率 $p(y_t \mid y_{<t},\vec{x})$ 时，将其分解为类别概率得分 $p(c_t \mid y_{<t},\vec{x})$ 与类别内部词汇概率得分 $p(y_t \mid c_t,\vec{y}_{<t},\vec{x})$ 的乘积，因而降低训练复杂度。本文使用四种词汇类别进行实验，分别为数字[NUM]、符号[SYM]、命名实体[NOUN]和其他词汇[UNK]，同时将词汇类别识别作为分类问题进行建模，训练分类器对 OOV 进行分类，具体实现在本文 4.2 节详述。

3.4.3 BPE 模型

参考(Sennrich et al., 2015)的工作，该模型对所有的 OOV 词汇进行处理，使用维吾尔语子词单元(subword units)表示方法，对低频词汇进行切分，由此增加稀疏词中子词的共现次数，有效解决数据稀疏问题。本文使用开源切分工具 subword-nmt（https://github.com/rsennrich/subword-nmt）对 OOV 词汇进行处理，并参考哈里旦木等[17]的工作，不对维吾尔语进行形态切分。

4　实验

4.1　实验设置

本文在汉语-维吾尔语统计机器翻译系统中进行实验，使用开源翻译平台 Moses（http://www.statmt.org/moses/）作为基线系统。训练数据来源于 2015 年 CWMT 公开的汉维新闻领域语料，共包含 11 万个平行双语句对，共有 64 851 个中文词汇以及 104 992 个维语词汇；开发集和测试集采用同领域数据，分别包含 1 095 个平行句对和 1 000 个平行句对。本文使用 SRILM[18]在训练数据上进行 5-gram 语言模型训练；使用斯坦福大学研发的分词器

（https://nlp.stanford.edu/software/segmenter.html）对汉语语句进行分词；使用 *grow-diag-final-and* 对齐策略，设置短语抽取最大长度为 8，共抽取出 4 934 572 个双语对齐短语；使用 MERT[19] 对 SMT 系统进行调参；使用大小写不敏感的 BLEU 值作为机器翻译性能的评价指标。

本文使用从 SMT 系统中抽取的双语对齐短语进行模型训练。首先，考虑到双语短语的语义匹配度，本文优先保留短语翻译概率最高的前 100 万个双语短语；然后，从中选择目标短语中各个词汇在源短语中至少含有一个对齐词汇的双语短语，在一定程度上保证内部词汇匹配度；最后，对于同一个源语言短语，保留较长的目标短语，以提高模型对于长短语的适应性。经过以上步骤，共筛选出约 40 万个对齐短语，作为本文中所提模型的训练数据。

4.2　模型训练

对于词汇类别分类器，本文使用 RNN 网络结构，设置神经网络隐藏层单元个数分别为 256-128-64-16-4，训练轮数为 200，初始学习率为 0.1，使用随机梯度下降(stochastic gradient descent, SGD)算法更新网络参数用于最小化损失函数，在输出层添加 *softmax* 激活函数输出四种词类的对应概率。当模型完成训练时，使用优化后的参数预测 OOV 的词类，并将最大值的所在类别对该词汇进行替换。

本模型中的编码器由正向和逆向 RNN 组成，各包含 100 个隐藏单元；解码器包含 100 个隐藏单元，使用带有 *maxout* 隐藏层的神经网络结构[20]；采用分批 SGD 算法结合学习率更新方法 Adadelta[21]训练本模型；设置分批训练数据规模为 100，模型训练轮数为 500。当训练完成时，使用优化后的网络参数评估双语短语的关联度得分，同时预测最有可能的目标短语。参考(Schwenk, 2012)中的思路，本文将统计翻译模型中的词汇化权重全部替换为双语关联度得分。

4.3　实验结果

本文对统计机器翻译系统和 NBROM 在句子级的机器翻译任务中的实验性能进行对比，同时考虑 OOV 处理策略、训练数据规模和目标词汇表大小，具体结果如表 1 所示。

表 1　机器翻译任务性能对比

Table 1　Experimental performance of machine translation tasks

模型	训练数据规模	词汇表大小	BLEU 值
Moses	110*k*	100*k*	38.16
NBROM + Unk			38.65 (+0.49)
NBROM + MultiClass	50*k*	30*k*	38.68 (+0.52)
NBROM + BPE			**38.75 (+0.59)**

由表 1 中数据可知，使用 NBROM 重新评估统计翻译模型中的词汇化权重，机器翻译性能有明显地提升；在训练数据规模为 50*k* 和词汇表大小为 30*k* 的条件下，BLEU 分值提升 0.49-0.59，证明本模型的有效性。其中 NBROM 结合 BPE 模型获得了本实验中最高的 BLEU 值 38.75；该方法将所有低频 OOV 词汇切分为字词形式，在一定程度上增加了稀疏词汇中字词的共现次数，减轻 OOV 词汇对于实验性能的影响，可以有效预测未登录词汇的生成概率，相比于基线系统提升 0.59，在本实验中性能提升最明显。对于 NBROM 相结合的 Unk 模型和 MultiClass 模型，后者稍优于前者的主要原因在于：Unk 模型对于所有的 OOV 词汇使用统一的符号[UNK]进行替换，赋予其相同权重；而 MultiClass 模型对词汇类别进行分类，在预测时充分考虑到词类信息，因而可以进一步提高 OOV 词汇的预测准确率。实验结果表明，基于双语短语的语义相关性和内部词汇匹配度等相关信息，本文提出的双语关联度优化模型可以在使用小规模的训练数据和词汇表的条件下，有效地提高汉语到维语的机器翻译任务性能。BPE 模型在机器翻译任务中的实验性能对比如表 2 所示。

表 2　BPE 模型在机器翻译任务中的实验性能对比

Table 2　Experimental performance of BPE model in machine translation task

训练数据规模	词汇表大小	BLEU 值
50*k*	30*k*	**38.75**
100*k*	40*k*	38.62
200*k*	50*k*	38.60

本文同样对 NBROM 结合 BPE 模型的实验性能进行了对比，并设置训练数据规模与词汇表大小成正比。由表 2 可知，BPE 模型在训练数据和词汇表规模较小的实验中性能最优，且随着训练数据和词汇表规模的扩大，实验性能降低，造成此结果的原因可能在于：维吾尔语的形态信息复杂并且词汇量巨大，在应用 BPE 模型时，需要将低频词汇切分为字词形式加入至目标词汇表中，在一定程度上影响维语语义信息的完整性，同时增加模型训练的复杂度，在预测 OOV 词汇的生成概率时面临数据稀疏性问题，因而减弱机器翻译任务的提升效果。

4.4　目标语言短语预测

本文对统计机器翻译系统和 NBROM 在短语级的机器翻译任务中的实验性能进行对比，测试数据为本文模型训练数据中随机抽取的 2 000 个双语对齐短语，并对测试集中的维语词汇量和短语中的平均词数进行统计，具体信息如表 3 所示。

表 3　短语生成任务性能对比

Table 3　Experimental performance of phrase generation task

模型	词汇量	平均词数	BLEU 值
Moses			91.27
NBROM + Unk	1,517	4.86	93.56 (+2.29)
NBROM + MultiClass			**93.76 (+2.49)**
NBROM + BPE			92.13 (+0.86)

表 4　维语短语预测实例

Table 4　Example of Uyghur phrase prediction

源语言短语	统计机器翻译系统	Score	NBROM + MultiClass	Rescore
金融 等 领域	مۇئامىلەقاتارلىقساھەلەردىكى پۇل	0.04556	ساھەلەردىكى مۇئامىلە قاتارلىق پۇل 领域 等 金融	0.79450
	قاتارلىقساھەلەردىكى پۇلمۇئامىلە	0.01479		
教育 、 文化 和	مائارىپ،مەدەنىيەتۋە	0.16721	، مەدەنىيەت ۋە مائارىپ 和 文化 、教育	0.83613
	مائارىپى،مەدەنىيەتۋە	0.03459		
活动 今天 启动 。	بۈگۈنچۈشتىنبۇرۇنباشلاندى كۆتۈرۈشپائالىيىتى	2.0e-15	بۈگۈن باشلاندى. پائالىيىتى 。启动 今天 活动	0.84382
	بۈگۈنچۈشتىنبۇرۇنباشلاندى پائالىيىتى	4.8e-11		
	پائالىيىتىبۈگۈنباشلاندى	0.02819		
	پائالىيەت،بۈگۈنباشلاندى	0.00101		

由表 3 中数据可知，使用 NBROM 预测源短语对应的目标短语，在准确性上明显高于统计机器翻译系统，证明了该方法在短语级机器翻译任务中的有效性。对于上述三种 OOV 词汇处理策略，MultiClass 模型在本实验的性能最好，相比于基线系统，BLEU 分值提升 2.49。NBROM 结合 BPE 模型的实验效果并不明显，造成此结果的原因可能是：使用 BPE 模型训练 NBROM 时，需要对维语词汇进行切分处理，故在生成目标词汇时引入过多的字词形式，降低预测短语中词汇的准确性以及完整性。此外，MultiClass 模型使用词汇类别

进行训练，相比于 Unk 模型，可进一步提高 OOV 词汇预测的准确率。

如上文中所述，NBROM 结合 MultiClass 模型可以预测最有可能的对齐目标短语，使之与源短语的匹配度最高；并可以重新评估词汇化权重，赋予双语短语更加合理的关联度得分。统计机器翻译系统与 NBROM 的目标语言短语预测实例如表 4 所示。其中 Score 表示统计翻译模型中的词汇化权重，Rescore 表示 NBROM 的双语关联度得分。统计机器翻译系统中相同的源语言短语对应多个目标短语，并保留相应的词汇化权重得分。由表 4 中数据可知，在语义内容以及词汇匹配度都较高的条件下，统计机器翻译系统中的词汇化权重分值较小，无法正确评估双语短语的对齐概率，与实际情况不符，因而降低翻译模型质量。与之相比，由于 NBROM 引入注意力机制，可以有效地捕获双语短语中的对齐词汇，因而可以合理地预测具有语义相关性和词汇匹配度的目标短语，同时赋予其相应的双语关联度得分，提高模型在短语级别的机器翻译任务中的实验性能。

5 结束语

针对汉维统计机器翻译系统中存在的语义无关性问题，本文提出了基于神经网络机器翻译方法的双语关联度优化模型，该模型引入注意力机制捕获双语短语的词对齐信息，并基于语义相关性和内部词汇匹配度重新评估双语短语的关联度得分，以此优化统计翻译模型中的词汇化权重；同时给定源短语，该模型可以预测匹配度最高的目标短语。实验结果表明，在使用较小规模的训练数据和词汇表的条件下，本文中提出的方法可以有效地提高短语级别和句子级别的机器翻译任务性能。

延续本文的研究方向，在后续工作中有以下思路：第一，由于词对齐结果中存在缺失、冗余、错误等问题，训练数据规模和词汇表大小会较大程度上地影响模型训练效果，因此考虑直接对词对齐结果进行优化；第二，本文只在汉维机器翻译任务中进行了数据分析和建模，对于其他语言对的翻译任务性能可能存在差异性，因此会在其他语言对上进行相关实验，提高模型的泛化能力；第三，对维语的词干词缀进行切分，以学习更多的词汇形态信息。

参考文献：

[1] Koehn P, Och F J, Marcu D. Statistical phrase-based translation [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics. 2003: 48-54.

[2] Och F J, Ney H. A systematic comparison of various statistical alignment models [J]. Computational linguistics, 2003, 29 (1): 19-51.

[3] 彭飞，吐尔根，艾山，等. 用于双语科技术语对齐的汉维文可比语料库构建 [J]. 新疆大学学报：自然科学版，2017，34 (3): 316-321. (Peng Fei, Tuergen Yibulayin, Aishan Wumaier, *et al.* 2017. Construction of Chinese-Uyghur comparable corpus for alignment of bilingual technical terms. Journal of Xinjiang University:Natural Science Edition , 34 (3): 316-321.)

[4] 米莉万•雪合来提，刘凯，吐尔根•依布拉音. 基于维吾尔语词干词缀粒度的汉维机器翻译 [J]. 中文信息学报，2015，29 (3): 201-206. (Miliwan Xuehelaiti, Liu Kai ,Turgun Ibrahim. 2015. Chinese-Uyghur machine translation model based on smallest translation units of stems and suffixes[J]. Journal of Chinese Information Processing, 29 (3): 201-206.)

[5] 潘一荣，李晓，杨雅婷，等. 面向汉维机器翻译的调序表重构模型 [J]. 计算机应用, 2018, 38 (5): 1283-1288. (Pan Yirong, Li Xiao, Yang Yating, *et al.* 2018. Reordering table reconstruction model for Chinese-Uyghur machine translation[J]. Journal of Computer Applications, 38 (5): 1283-1288.)

[6] Schwenk H. Continuous space translation models for phrase-based statistical machine translation [J]. Proceedings of COLING 2012: Posters, 2012: 1071-1080.

[7] Son L H, Allauzen A, Yvon F. Continuous space translation models with neural networks [C]// Proc of Conference of the North American Chapter of the Association for Computational linguistics: Human language technologies. Association for Computational Linguistics, 2012: 39-48.

[8] Zou W Y, Socher R, Cer D, *et al.* Bilingual word embeddings for phrase-based machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 1393-1398.

[9] Cho K, Van Merriënboer B, Gulcehre C, *et al.* 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 1406. 1078.

[10] Bahdanau D , Cho K , Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409. 0473.

[11] Sennrich R , Haddow B , Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv: 1508. 07909.

[12] Papineni K, Roukos S, Ward T, *et al.* BLEU: a method for automatic evaluation of machine translation [C]// Proc of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.2002: 311-318.

[13] Och F J, Ney H. Giza+: training of statistical translation models[Z]. 2000.

[14] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Trans on Signal Processing, 1997, 45 (11): 2673-2681.

[15] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in neural information processing systems. 2014: 3104-3112.

[16] Jean S. , Cho K. , Memisevic R. , et al. 2014. On using very large target vocabulary for neural machine translation[J]. arXiv preprint arXiv: 1412. 2007.

[17] 哈里旦木•阿布都克里木，刘洋，孙茂松. 神经机器翻译系统在维吾尔语-汉语翻译中的性能对比 [J]. 清华大学学报:自然科学版 , 2017, 57 (1): 1-6. (Halidanmu Abudukelimu, Liu Yang, Sun Maosong. Performance comparison of neural machine translation systems in Uyghur-Chinese translation[J]. Journal of Tsinghua University :Science and Technology , 57 (1): 1-6.)

[18] Stolcke, A. SRILM: an extensible language modeling toolkit[C]//Proc of the 7th International Conference on Spoken Language Processing,volume 2.2002:901-904.

[19] Och F J. Minimum error rate training in statistical machine translation [C]// Proc of the 41st Annual Meeting on Association for Computational Linguistics,volume 1. 2003: 160-167.

[20] Goodfellow I J, Warde-Farley D, Mirza M, *et al.* Maxout networks[J]. 2013.arXiv preprint arXiv: 1302. 4389.

[21] Zeiler M D. ADADELTA: an adaptive learning rate method[J]. 2012.arXiv preprint arXiv: 1212. 5701.